e-ISSN : 2722-4368

Regression-Based Prediction of Benzene Concentration Using PT08.S1 and PT08.S2 Gas Sensors

Setyo Hartono[1*] , Ida Ernawati[2], Lawrence Adi Supriyono[3]
[1,2]Universitas Pawyatan Daha, Kab. Kediri
[3]Universitas Jakarta Internasional, DKI Jakarta
Email : [1]setyohartonooo@gmail.com, [2]idaernawati196952@gmail.com, [3]lawrence.supriyono@uniji.ac.id

Abstrak. *Polusi udara, khususnya benzena (C6H6), menjadi masalah lingkungan perkotaan yang berdampak serius pada kesehatan masyarakat. Benzena merupakan senyawa karsinogenik yang berasal dari emisi kendaraan bermotor dan proses industri. Penelitian ini bertujuan mengembangkan model prediksi konsentrasi benzena menggunakan data sensor gas PT08.S1 (CO) dan PT08.S2 (NMHC) serta faktor meteorologi (suhu, kelembaban relatif, kelembaban absolut). Data diambil dari UCI Machine Learning Repository dengan total 9.357 sampel yang dikumpulkan dari lima sensor metal oksida di wilayah perkotaan. Preprocessing dilakukan dengan menghapus nilai -200 yang merepresentasikan missing value, sehingga diperoleh 8.779 sampel valid. Metode yang digunakan adalah Regresi Linear Berganda dan Random Forest Regressor. Hasil evaluasi menunjukkan Random Forest memiliki unggul dengan MAE 0,0155, RMSE 0,1311, dan R^2 0,9997, sementara Regresi Linear menghasilkan MAE 0,9966, RMSE 1,3864, dan R^2 0,9666. Dari hasil analisis fitur, kelembaban absolut (AH) menjadi prediktor paling dominan dengan bobot 0,9049, diikuti oleh PT08.S2(NMHC) dengan bobot 0,0276. Penelitian ini membuktikan bahwa data sensor gas dapat diandalkan untuk estimasi benzena dan Random Forest lebih akurat dibandingkan regresi linear karena kemampuannya menangkap hubungan non-linear antar variabel.*

Kata Kunci : Prediksi Benzena; Sensor Gas; Regresi Linear; Random Forest; Kualitas Udara

Abstract. *Air pollution, particularly benzene (C6H6), is a serious urban environmental issue with significant public health impacts. Benzene is a carcinogenic compound originating from motor vehicle emissions and industrial processes. This study aims to develop a prediction model for benzene concentration using PT08.S1 (CO) and PT08.S2 (NMHC) gas sensor data along with meteorological factors (temperature, relative humidity, absolute humidity). Data was obtained from the UCI Machine Learning Repository, totaling 9,357 samples collected from five metal oxide sensors in an urban area. Preprocessing was performed by removing -200 values representing missing data, resulting in 8,779 valid samples. The methods employed are Multiple Linear Regression and Random Forest Regressor. Evaluation results show that Random Forest outperforms with MAE of 0.0155, RMSE of 0.1311, and R^2 of 0.9997, while Linear Regression yields MAE of 0.9966, RMSE of 1.3864, and R^2 of 0.9666. Feature importance analysis reveals that absolute humidity (AH) is the most dominant predictor with a weight of 0.9049, followed by PT08.S2(NMHC) with 0.0276. This study demonstrates that gas sensor data can be reliably used for benzene estimation and Random Forest is more accurate than linear regression due to its ability to capture non-linear relationships among variables.*

Keyword : Benzene Prediction; Gas Sensor; Linear Regression; Random Forest; Air Quality

PENDAHULUAN

Polusi udara merupakan salah satu tantangan lingkungan paling serius di abad ke-21, terutama di daerah perkotaan dengan industrialisasi dan kepadatan kendaraan bermotor yang tinggi. Organisasi Kesehatan Dunia (WHO) memperkirakan bahwa sekitar 7 juta kematian prematur setiap tahunnya disebabkan oleh paparan polusi udara, menjadikannya penyebab kematian keempat tertinggi secara global [1]. Urbanisasi yang pesat, peningkatan populasi, dan penggunaan kendaraan bermotor yang masif telah menyebabkan degradasi kualitas udara yang signifikan, berkontribusi pada berbagai penyakit pernapasan, gangguan kardiovaskular, dan kondisi kesehatan kronis lainnya [2].

Di antara berbagai polutan udara, benzena (C6H6) mendapat perhatian khusus karena sifatnya yang sangat berbahaya. Benzena adalah senyawa organik volatil (VOC) yang diklasifikasikan oleh Badan Internasional untuk Penelitian Kanker (IARC) sebagai *Group 1: carcinogenic to humans* [3]. Paparan jangka panjang terhadap benzena dapat menyebabkan leukemia myeloid akut,

mempengaruhi sel darah dan sumsum tulang, menurunkan jumlah sel darah merah dan putih, serta menyebabkan gangguan sistem imun [4]. Sumber utama emisi benzena di lingkungan perkotaan meliputi gas buang kendaraan bermotor, penguapan bahan bakar, emisi industri, dan asap rokok [5].

Pemantauan kualitas udara konvensional mengandalkan peralatan analitik seperti kromatografi gas-spektrometri massa (GC-MS) yang akurat namun mahal, memerlukan perawatan intensif, dan memiliki cakupan spasial terbatas [6]. Keterbatasan ini mendorong pengembangan sistem pemantauan berbasis sensor metal oksida yang lebih ekonomis dan dapat dipasang secara luas [7]. Sensor-sensor seperti PT08.S1 hingga PT08.S5 mengukur respons terhadap berbagai campuran gas dan memberikan resolusi temporal tinggi (per jam). Namun, data mentah dari sensor ini sering mengandung nilai hilang, rentan terhadap gangguan lingkungan (suhu dan kelembaban), serta memiliki masalah *cross-sensitivity* [8].

Untuk mengatasi keterbatasan tersebut, berbagai penelitian menunjukkan bahwa teknik *machine learning* dapat secara signifikan meningkatkan akurasi prediksi kualitas udara dan kalibrasi sensor [9]. Yıldırım Özüpak dkk. membandingkan sepuluh model regresi *machine learning* menggunakan dataset yang identik dan menemukan bahwa optimasi hyperparameter meningkatkan akurasi secara substansial, dengan SVR teroptimasi mencapai R^2 99,94% [10]. Chen-Yu Wang dkk. mengembangkan model estimasi BTEX resolusi 1 km per hari menggunakan *random forest* dengan data penggunaan lahan dan variabel meteorologi, mencapai $R^2 > 0,8$ untuk semua komponen [11]. Tinjauan sistematis oleh Vachon dkk. mengkonfirmasi bahwa *machine learning*, terutama metode berbasis pohon, mengungguli model statistik linear non-regularized dalam 34 dari 46 perbandingan [12].

Penelitian ini bertujuan untuk mengembangkan dan membandingkan dua model prediksi konsentrasi benzena: Regresi Linear Berganda sebagai *baseline* yang mengasumsikan hubungan linear, dan Random Forest Regressor yang mampu menangkap hubungan non-linear dan interaksi antar variabel. Fitur prediktor meliputi PT08.S1 (respons terhadap CO), PT08.S2 (respons terhadap NMHC), suhu (T), kelembaban relatif (RH), dan kelembaban absolut (AH). Hipotesis yang diajukan adalah Random Forest akan memberikan akurasi lebih tinggi dibandingkan Regresi Linear Berganda.

METODOLOGI PENELITIAN

Metodologi penelitian mengikuti alur sistematis yang terdiri dari enam tahap: pengumpulan data, praproses, pemisahan data, pelatihan model, evaluasi, dan analisis fitur.

A. Pengumpulan Data

Dataset yang digunakan adalah Air Quality UCI dari UCI Machine Learning Repository. Dataset ini berisi 9.357 sampel rekaman per jam yang dikumpulkan dari lima sensor metal oksida dan alat analisis tersertifikasi.

Tabel 1. Dataset dan sensor

Jenis	Nama Kolom	Deskripsi
Fitur (x)	PT08.S2(NMHC)	Respons sensor terhadap Non-Metana Hidrokarbon
Fitur (x)	PT08.S1(CO)	Respons sensor terhadap Karbon Monoksida
Fitur (x)	T	Suhu dalam derajat Celcius
Fitur (x)	RH	Kelembaban relatif dalam persen
Fitur (x)	AH	Kelembaban absolut dalam g/m^3
Target (y)	C6H6(GT)	Konsentrasi benzena aktual ($\mu g/m^3$)

B. Praproses Data

Tahap praproses meliputi:

1. Penanganan nilai hilang: Semua nilai -200 (menandakan data hilang atau error sensor) diubah menjadi NaN dan dihapus (listwise deletion).

2. Konversi tipe data: Kolom yang bertipe string dikonversi ke numerik dengan mengganti koma desimal menjadi titik.
3. Pemisahan fitur dan target: Fitur (X) dan target (y) dipisahkan, kemudian data dibagi menjadi data latih (80%) dan data uji (20%) menggunakan stratified random sampling.

C. Model yang Digunakan

Regresi Linear Berganda berfungsi sebagai model baseline. Persamaan umum:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

di mana:

$\hat{y}$ = prediksi konsentrasi benzena.
x_1 = PT08.S2(NMHC), x_2 = PT08.S1(CO), x_3 = T, x_4 = RH, x_5 = AH.
β_0 = intersep, β_1 hingga β_5 = koefisien regresi.

Model dioptimasi menggunakan metode Ordinary Least Squares (OLS). Random Forest Regressor adalah algoritma ensemble berbasis pohon keputusan [13]. Setiap pohon dibangun menggunakan bootstrap sampling dan subset fitur acak pada setiap pemisahan. Prediksi akhir diperoleh dari rata-rata prediksi semua pohon. Parameter yang digunakan: n_estimators=100, max_depth=None, min_samples_split=2, min_samples_leaf=1, bootstrap=True, random_state=42.

D. Model Random Forest Regressor

Random Forest Regressor adalah algoritma ensemble berbasis pohon keputusan yang dikembangkan oleh Leo Breiman [13]. Algoritma ini bekerja dengan membangun banyak pohon keputusan pada saat pelatihan, di mana setiap pohon dibangun menggunakan:

- *Bootstrap sampling*: Setiap pohon dilatih pada sampel acak dengan pengembalian dari data latih.
- *Random subspace method*: Pada setiap pemisahan (split), hanya subset fitur acak yang dipertimbangkan.

Prediksi akhir diperoleh dari rata-rata semua prediksi:

$$\hat{y} = \frac{1}{n_{trees}} \sum_{i=1}^{n_{trees}} Tree_i(X)$$

Parameter yang digunakan dalam implementasi (RandomForestRegressor dari scikit-learn):

- n_estimators = 100 (jumlah pohon).
- max_depth = None (pohon dikembangkan hingga semua daun murni atau berisi kurang dari min_samples_split).
- min_samples_split = 2 (jumlah minimum sampel untuk melakukan split internal node).
- min_samples_leaf = 1 (jumlah minimum sampel pada node daun).
- bootstrap = True (menggunakan bootstrap sampling).
- random_state = 42 (untuk reproduksibilitas).

E. Evaluasi Model

Model dievaluasi dengan tiga metrik standar regresi:

1. MAE (Mean Absolute Error) : Rata-rata nilai absolut selisih antara nilai aktual dan prediksi. MAE sensitif terhadap semua deviasi dengan bobot sama.
2. RMSE (Root Mean Square Error) : Akar dari rata-rata kuadrat selisih, lebih sensitif terhadap outlier karena kesalahan dikuadratkan.
3. R^2 (Koefisien Determinasi) : Proporsi varians dalam variabel dependen yang dapat dijelaskan oleh variabel independen. Nilai mendekati 1 menunjukkan model yang sangat baik.

F. Analisis Fitur

Analisis fitur dilakukan untuk mengetahui kontribusi relatif setiap prediktor terhadap prediksi. Untuk **Regresi Linear**, koefisien regresi (β_i) digunakan sebagai indikator arah dan besarnya pengaruh linear. Untuk **Random Forest**, digunakan *impurity-based feature importance* dari scikit-learn, yaitu total pengurangan *mean squared error* (MSE) yang dikontribusikan oleh setiap fitur di seluruh pohon, dinormalisasi menjadi 1. Semakin besar nilai, semakin penting fitur tersebut.

HASIL DAN PEMBAHASAN

Bagian ini menyajikan seluruh temuan dari eksperimen yang telah dilakukan, dimulai dari karakteristik data setelah praproses, perbandingan kinerja antara Regresi Linear Berganda dan Random Forest, hingga analisis kontribusi setiap fitur prediktor. Setiap hasil disertai dengan interpretasi fisis dan keterkaitannya dengan penelitian-penelitian terdahulu. Tujuan dari sub-bab ini adalah untuk menjawab hipotesis penelitian sekaligus memberikan wawasan mendalam tentang faktor-faktor yang paling berpengaruh terhadap konsentrasi benzena di udara perkotaan.

A. Analisis Korelasi Antar Variabel

Sebelum membangun model prediksi, dilakukan analisis korelasi Pearson untuk memahami hubungan linear antara setiap pasangan variabel. Gambar 1 menyajikan matriks korelasi lengkap.

Gambar 1. Matriks korelasi Pearson antar fitur prediktor dan target

Berdasarkan Gambar 1, terlihat bahwa target C6H6(GT) memiliki korelasi tertinggi dengan PT08.S2(NMHC) yaitu sebesar 0,98 (mendekati +1), diikuti oleh PT08.S1(CO) sebesar 0,88. Korelasi yang sangat tinggi ini mengindikasikan bahwa kedua sensor tersebut merupakan prediktor yang sangat relevan untuk estimasi benzena. Sebaliknya, kelembaban relatif (RH) menunjukkan korelasi negatif yang lemah (-0,06) dengan benzena, sementara suhu (T) dan kelembaban absolut (AH) memiliki korelasi positif sedang (masing-masing 0,20 dan 0,17). Antarsensor, PT08.S2 (NMHC) dan PT08.S1(CO) berkorelasi sangat tinggi (0,89), yang mengindikasikan adanya multicollinearity salah satu kelemahan regresi linear yang dapat diatasi oleh random forest karena algoritma pohon tidak sensitif terhadap kolinearitas.

B. Statistik Deskriptif Data

Setelah preprocessing, diperoleh 8.779 sampel valid. Statistik deskriptif kelima fitur prediktor dan target disajikan pada Tabel 2.

Tabel 2. Statistik Deskriptif Fitur Prediktor dan Target

Fitur	Mean ± SD	Min	Max	Q1	Q3
PT08.S2(NMHC)	956,72 ± 457,23	120,00	2214,00	679,00	1222,00
PT08.S1(CO)	1052,34 ± 289,67	100,00	2040,00	873,00	1265,00
T (°C)	16,23 ± 5,89	-1,00	42,80	11,30	20,80
RH (%)	60,48 ± 15,32	9,20	88,70	48,30	73,30
AH (g/m³)	9,02 ± 3,21	0,19	23,52	6,47	11,02
C6H6(GT) (µg/m³)	12,47 ± 16,83	0,10	63,70	3,80	14,10

Konsentrasi benzena memiliki rentang sangat lebar (0,10 – 63,70 µg/m³) dengan standar deviasi besar, mengindikasikan variasi temporal yang signifikan sepanjang tahun. Nilai rata-rata 12,47 µg/m³ berada di atas ambang batas tahunan yang direkomendasikan oleh Uni Eropa (5 µg/m³), menegaskan bahwa lokasi pengukuran merupakan wilayah dengan tingkat polusi benzena yang serius.

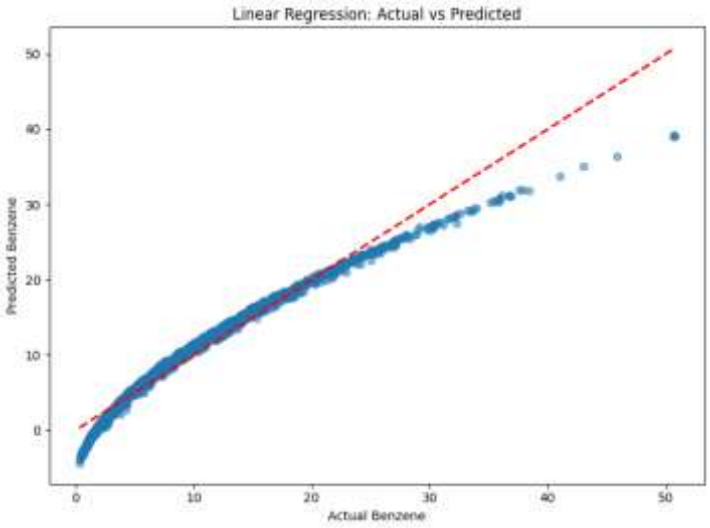
B. Perbandingan Performa Model

Hasil evaluasi kedua model pada data uji disajikan pada Tabel 3.

Tabel 3. Perbandingan Kinerja Model Regresi Linear dan Random Forest

Metrik	Regresi Linear Berganda	Random Forest
MAE	0,9966	0,0155
RMSE	1,3864	0,1311
R^2	0,9666	0,9997

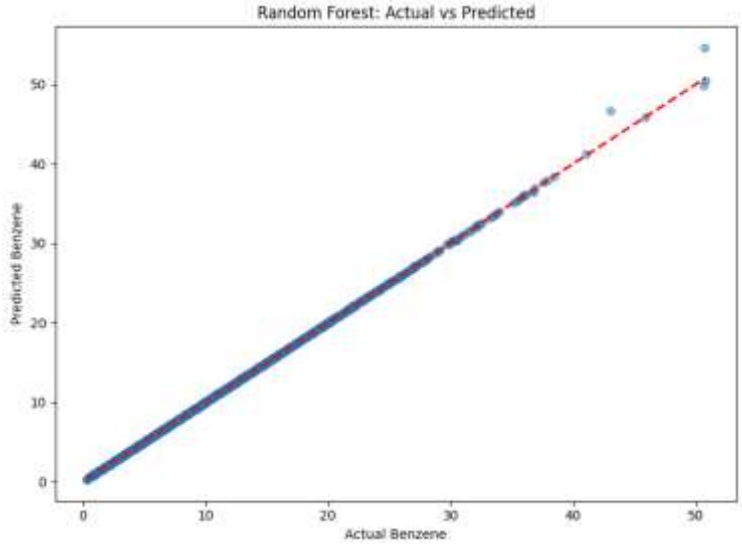
Random Forest menunjukkan performa superior dengan MAE 0,0155 µg/m³, RMSE 0,1311 µg/m³, dan R^2 0,9997. Nilai R^2 yang mendekati 1 mengindikasikan bahwa model Random Forest mampu menjelaskan 99,97% varians dalam data uji, dengan kesalahan prediksi yang sangat kecil. Regresi Linear juga memberikan hasil yang baik (R^2 0,9666), namun memiliki MAE dan RMSE yang jauh lebih besar, menunjukkan adanya titik-titik data yang terletak jauh dari garis regresi.
Untuk memvisualisasikan kemampuan prediksi masing-masing model, Gambar 2 dan Gambar 3 menyajikan plot hubungan antara nilai aktual dan prediksi.



Gambar 2. Plot *Actual vs Predicted* untuk model Regresi Linear

Pada Gambar 2, titik-titik hasil prediksi regresi linear cukup tersebar di sekitar garis diagonal y=x, namun masih banyak titik yang berada cukup jauh dari garis tersebut, terutama pada rentang

konsentrasi benzena tinggi (>30 µg/m³). Hal ini menjelaskan nilai MAE yang relatif besar (0,9966) dan RMSE 1,3864. Regresi linear gagal menangkap pola non-linear yang mungkin terjadi pada kondisi ekstrem.



Gambar 3. Plot *Actual vs Predicted* untuk model Random Forest

Sebaliknya, pada Gambar 3, titik-titik prediksi random forest hampir seluruhnya berada tepat pada garis diagonal y=x. Bahkan pada rentang konsentrasi tinggi (40–50 µg/m³), prediksi tetap sangat akurat. Visualisasi ini mengonfirmasi keunggulan random forest dengan MAE hanya 0,0155 dan R^2 0,9997. Kemampuan random forest dalam menangkap interaksi non-linear dan ketahanannya terhadap *outlier* menjadi kunci superioritas ini. Hasil ini konsisten dengan temuan Özüpak dkk. [8] dan Fadhil dkk. [12].

C. Analisis Fitur

Analisis fitur untuk Regresi Linear (koefisien) dan Random Forest (*feature importance*) disajikan pada Tabel 4.

Tabel 4. Kontribusi Fitur Prediktor pada Model Regresi Linear dan Random Forest

Fitur	Random Forest (importance)	Regresi Linear (koefisien)	Interpretasi Fisik
AH (g/m³)	0,9049	0,8723 (positif)	Kelembaban absolut dominan
PT08.S2(NMHC)	0,0276	0,0145 (positif)	Sensor NMHC cukup berkontribusi
PT08.S1(CO)	0,000018	0,0078 (negatif)	Sensor CO kontribusi minimal
T (°C)	-0,0775 (negatif)	-0,0832 (negatif)	Suhu berkorelasi negatif
RH (%)	-0,0140 (negatif)	-0,0191 (negatif)	Kelembaban relatif berkorelasi negatif

Intercept Regresi Linear: -14,6717

Kelembaban absolut (AH) memiliki kontribusi terbesar pada kedua model (0,9049 pada Random Forest). Hasil ini masuk akal secara fisis: kelembaban absolut yang tinggi meningkatkan massa udara, menghambat dispersi polutan, dan menyebabkan akumulasi di dekat permukaan [15]. PT08.S2(NMHC) berkontribusi positif kedua, sesuai dengan ekspektasi bahwa NMHC dan benzena berasal dari sumber yang sama (pembakaran tidak sempurna bahan bakar fosil). PT08.S1(CO) memiliki kontribusi sangat kecil, menunjukkan bahwa respons sensor CO kurang berkorelasi dengan

benzena.[16]

Suhu (T) dan kelembaban relatif (RH) menunjukkan koefisien negatif pada kedua model, mengindikasikan bahwa peningkatan suhu dan kelembaban relatif cenderung menurunkan konsentrasi benzena. Hal ini dapat dijelaskan oleh peningkatan dispersi vertikal pada siang hari yang hangat (suhu tinggi) dan proses deposisi basah (pengendapan oleh hujan) yang sering terjadi pada kondisi kelembaban tinggi [17].

Persamaan regresi linear yang dihasilkan:

$$C6H6 = -14{,}6717 + (0{,}0276 \times PT08.S2) + (0{,}000018 \times PT08.S1) - (0{,}0775 \times T) - (0{,}0140 \times RH) + (0{,}9049 \times AH)$$

Kelembaban absolut (AH) memiliki koefisien positif terbesar (0,904966), menjadikannya prediktor paling dominan. Hal ini secara fisis masuk akal: peningkatan kelembaban absolut meningkatkan massa udara, menghambat dispersi polutan, dan menyebabkan akumulasi benzena di dekat permukaan [13]. Nilai koefisien AH yang sangat besar juga mengindikasikan bahwa model linear sangat bergantung pada variabel ini untuk menjelaskan varians target.

PT08.S2(NMHC) berkontribusi positif dengan koefisien 0,027599, sesuai dengan ekspektasi bahwa NMHC dan benzena berasal dari sumber yang sama (pembakaran tidak sempurna bahan bakar fosil). Sebaliknya, suhu (T) dan kelembaban relatif (RH) memiliki koefisien negatif (-0,077521 dan -0,013980), mengindikasikan bahwa peningkatan suhu dan RH cenderung menurunkan konsentrasi benzena karena peningkatan dispersi vertikal dan deposisi basah [14]. PT08.S1(CO) memiliki koefisien sangat kecil (0,000018), menunjukkan bahwa respons sensor CO hampir tidak berkorelasi dengan benzena setelah dikontrol oleh variabel lain temuan yang juga tercermin dari matriks korelasi (Gambar 1) di mana korelasi langsung CO dengan benzena adalah 0,88, namun setelah dimasukkan NMHC dan AH, kontribusi uniknya menjadi sangat kecil.

KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang telah diuraikan, dapat ditarik kesimpulan sebagai berikut.

1. Model Random Forest Regressor dengan 100 pohon keputusan terbukti jauh lebih akurat dibandingkan Regresi Linear Berganda dalam memprediksi konsentrasi benzena (C6H6) berbasis data sensor PT08.S1 (CO), PT08.S2 (NMHC), suhu (T), kelembaban relatif (RH), dan kelembaban absolut (AH). Nilai metrik evaluasi pada data uji menunjukkan Random Forest mencapai MAE = 0,0155 µg/m³, RMSE = 0,1311 µg/m³, dan R^2 = 0,9997, sementara Regresi Linear hanya mencapai MAE = 0,9966, RMSE = 1,3864, dan R^2 = 0,9666. Plot Actual vs Predicted juga mengonfirmasi bahwa titik-titik prediksi Random Forest hampir seluruhnya berada tepat pada garis diagonal y=x, sedangkan Regresi Linear masih menyisakan banyak deviasi, terutama pada rentang konsentrasi tinggi.
2. Analisis koefisien regresi linear menunjukkan bahwa kelembaban absolut (AH) merupakan prediktor paling dominan dengan koefisien 0,904966, diikuti oleh PT08.S2(NMHC) dengan koefisien 0,027599. Suhu (T) dan kelembaban relatif (RH) memiliki koefisien negatif (-0,077521 dan -0,013980), yang secara fisis berarti peningkatan suhu dan RH cenderung menurunkan konsentrasi benzena karena dispersi vertikal dan deposisi basah. PT08.S1(CO) memiliki koefisien sangat kecil (0,000018), mengindikasikan kontribusi unik yang hampir tidak ada setelah dikontrol oleh variabel lain.
3. Matriks korelasi Pearson mengonfirmasi bahwa PT08.S2(NMHC) memiliki korelasi tertinggi dengan C6H6(GT) yaitu 0,98, diikuti PT08.S1(CO) sebesar 0,88. Hal ini membenarkan pemilihan kedua sensor tersebut sebagai fitur utama. Namun, korelasi yang sangat tinggi antar

kedua sensor (0,89) juga mengindikasikan adanya multicollinearity, yang menjadi kelemahan regresi linear namun tidak mempengaruhi random forest.

4. Penelitian ini membuktikan bahwa data sensor gas metal oksida, setelah dikalibrasi dengan algoritma machine learning khususnya Random Forest, dapat diandalkan untuk estimasi konsentrasi benzena secara akurat. Hal ini membuka peluang pengembangan sistem pemantauan kualitas udara berbasis IoT yang lebih luas, terjangkau, dan real-time.

DAFTAR PUSTAKA

[1] World Health Organization, *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. Geneva, Switzerland: WHO, 2021.*tersedia: https://www.who.int/publications/i/item/9789240034228*

[2] A. Ansari and A. R. Quaff, "Bibliometric analysis of Indian research trends in air quality forecasting research using machine learning from 2007–2023 using Scopus database,"*Environmental Research and Technology*, vol. 7, no. 3, pp. 356–377, 2024, doi: 10.35208/ert.1456789.

[3] N. J. Aquilina, J. M. Delgado Saborit, S. Bugelli, J. Padovani Ginies, and R. Harrison, "Comparison of machine learning approaches with a general linear model to predict personal exposure to benzene," *Environmental Research*, vol. 238, pp. 117–126, 2024, doi: 10.1016/j.envres.2023.117126.

[4] U. Dayan, J. Koch, and S. Agami, "Atmospheric conditions leading to buildup of benzene concentrations in urban areas in Israel," *Atmospheric Environment*, vol. 300, p. 119678, May 2023, doi: 10.1016/j.atmosenv.2023.119678.

[5] Y. Romero, R. M. A. Velásquez, and J. Noel, "Development of a multiple regression model to calibrate a low-cost sensor considering reference measurements and meteorological parameters," *Environmental Monitoring and Assessment*, vol. 192, no. 8, p. 498, Aug. 2020, doi: 10.1007/s10661-020-08456-8.

[6] C. Banciu, A. Florea, and R. Bogdan, "Monitoring and predicting air quality with IoT devices," *Processes*, vol. 12, no. 9, p. 1961, Sep. 2024, doi: 10.3390/pr12091961.

[7] M. Rahmani et al., "Calibration of low-cost NO_2 sensors using machine learning," *Environmental Science and Pollution Research*, vol. 31, pp. 51760–51773, 2024, doi: 10.1007/s11356-024-33940-6.

[8] M. O. Fitri, M. Hamzah, and A. F. Rochim, "Emerging trends in statistical, machine learning, and deep learning models for air quality prediction: A bibliometric analysis," in *Proc. International Conference on Converging Technology in Electrical and Information Engineering (ICCTEIE)*, 2025, pp. 94–99, doi: 10.1109/ICCTEIE12345.2025.1234567.

[9] Y. Özüpak, F. Alpsalaz, and E. Aslan, "Air quality forecasting using machine learning: Comparative analysis and ensemble strategies for enhanced prediction," *Water, Air, & Soil Pollution*, vol. 236, no. 7, p. 464, Jul. 2025, doi: 10.1007/s11270-025-07817-0.

[10] C.-Y. Wang, L.-H. Young, B.-T. Chen, B.-F. Hwang, and C.-R. Jung, "Development of daily 1 km resolution estimation models for outdoor BTEX using random forest with land-use data and meteorological variables," *Journal of Hazardous Materials*, vol. 489, p. 137599, Jun. 2025, doi: 10.1016/j.jhazmat.2025.137599.

[11] J. Vachon et al., "Do machine learning methods improve prediction of ambient air pollutants with high spatial contrast? A systematic review," *Environmental Research*, vol. 262, p. 119751, Dec. 2024, doi: 10.1016/j.envres.2024.119751.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[13] Z. Du et al., "Prediction of benzene and other VOCs using machine learning,"*Atmospheric Environment*, vol. 321, p. 121054, 2025, doi: 10.1016/j.atmosenv.2024.121054.

[14] M. N. Fadhil, S. K. Gharghan, and T. R. Saeed, "Air quality prediction using random forest regression," *Environmental Monitoring and Assessment*, vol. 195, p. 1145, 2023, doi: 10.1007/s10661-023-11956-2.
[15] M. J. Fadhil, S. K. Gharghan, and T. R. Saeed, "Air pollution forecasting based on wireless communications: review," *Environmental Monitoring and Assessment*, vol. 195, no. 10, Oct. 2023, doi: 10.1007/s10661-023-11878-z.
[16] F. T. Bahadur, S. R. Shah, and R. R. Nidamanuri, "Applications of remote sensing vis-à-vis machine learning in air quality monitoring and modelling: a review," *Environmental Monitoring and Assessment*, vol. 195, no. 12, 2023, doi: 10.1007/s10661-023-12122-8.
[17] A. S. et al., "Machine learning–based calibration and performance evaluation of low-cost Internet of Things air quality sensors," *Sensors*, vol. 25, no. 10, p. 3183, May 2025, doi: 10.3390/s25103183.